

Creating a network of networks in human genome epidemiology

John P.A. Ioannidis, MD

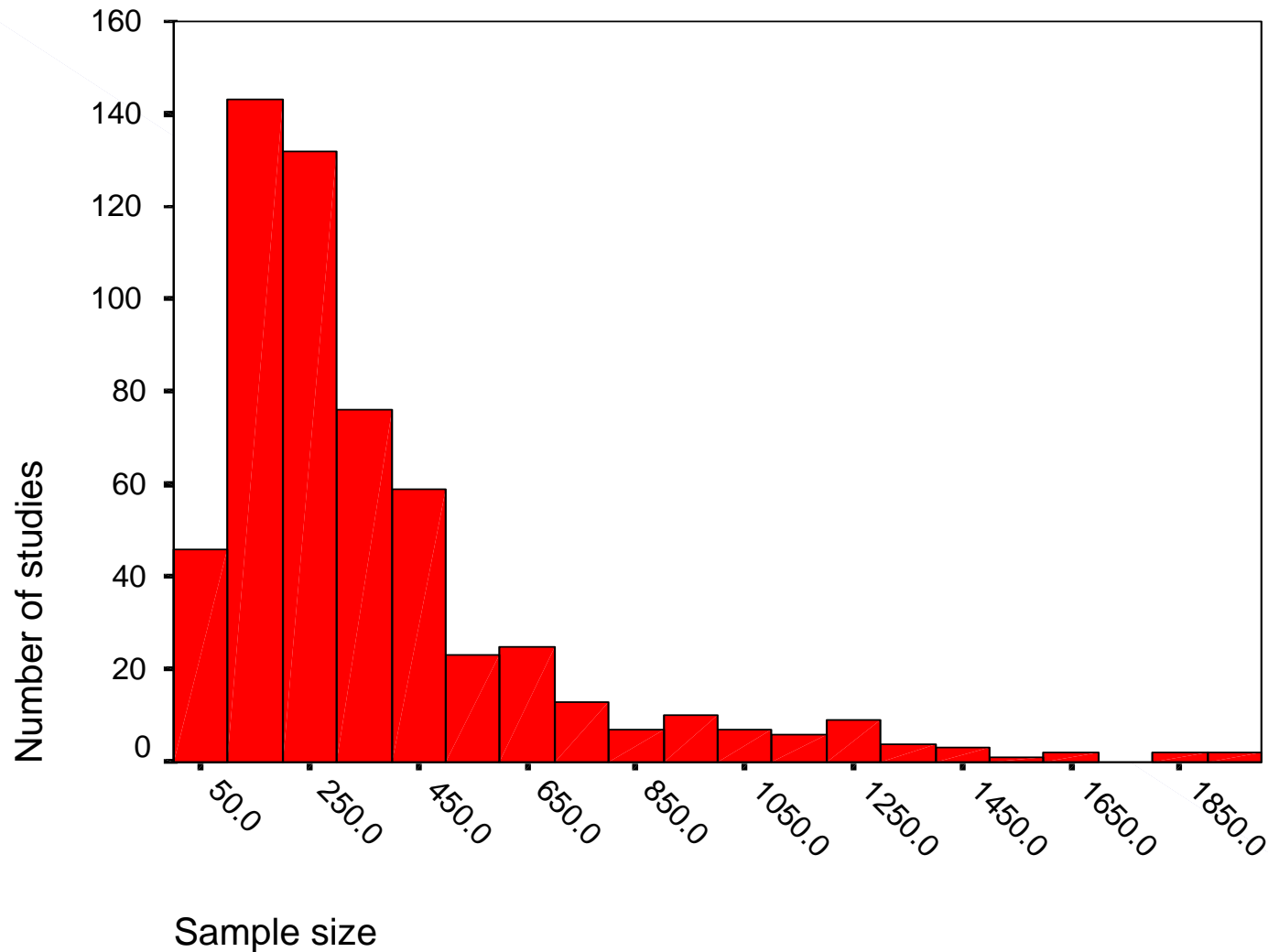
International Biobank and Cohort Studies meeting

Atlanta Feb 7-8, 2005

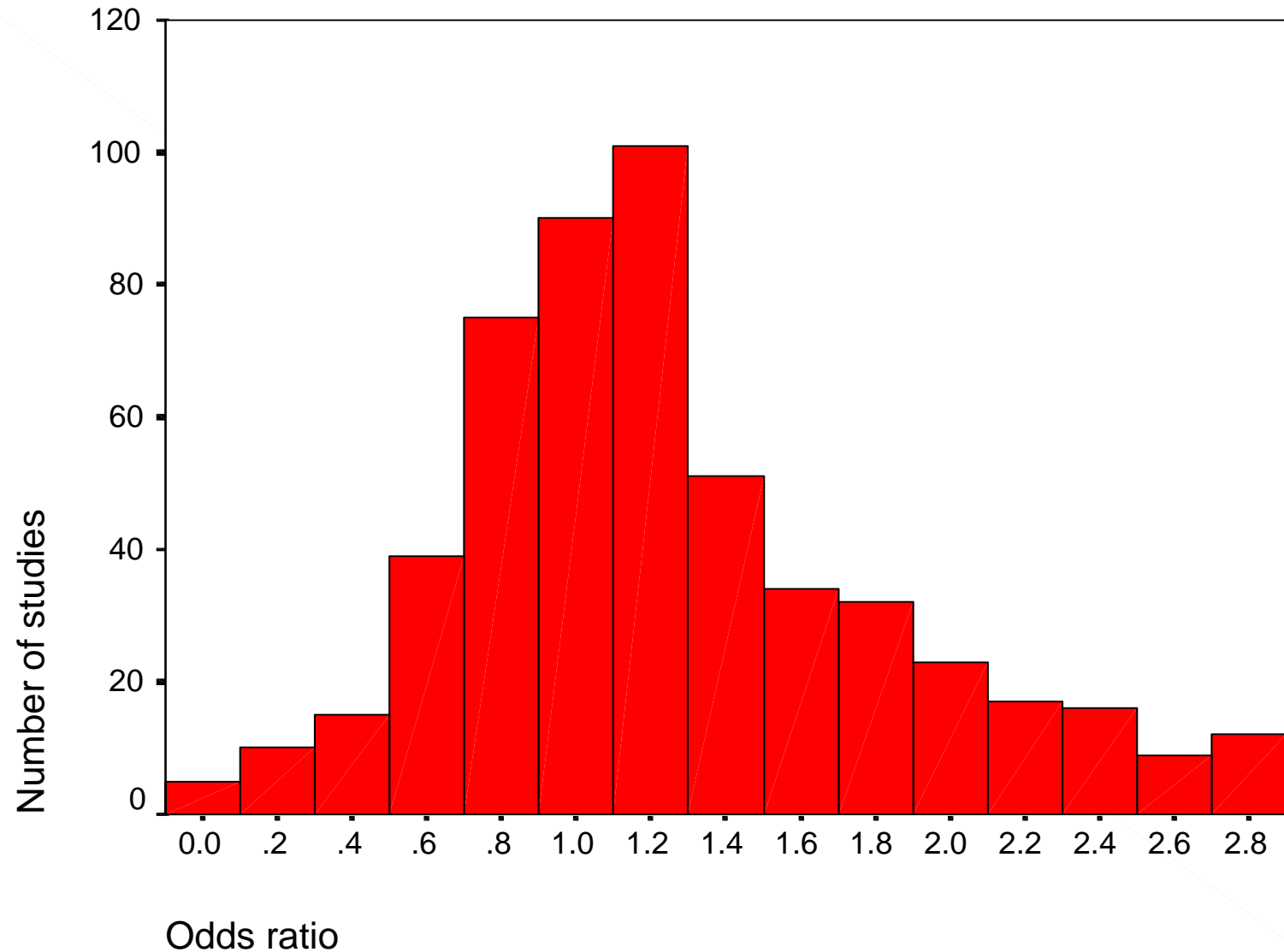
Empirical evidence on problems and biases in genetic epidemiology

- Small studies and small effects
- Multiplicity of analyses for small effects
- Shaky foundations of biological plausibility
- Different results in early vs. late studies
- Spuriously clear genetic (or other biological) contrasts
- Large vs. small studies
- Proteus phenomenon (alternating extreme effects)
- Racial and other subgroup effects
- Language bias and reverse language bias
- Available, hidden, and unavailable evidence
- Standardization issues for polymorphic markers, qualitative traits, intermediate endpoints, etc.
- Too much analytical liberty

Small sample size of individual studies



Small effect sizes in individual studies



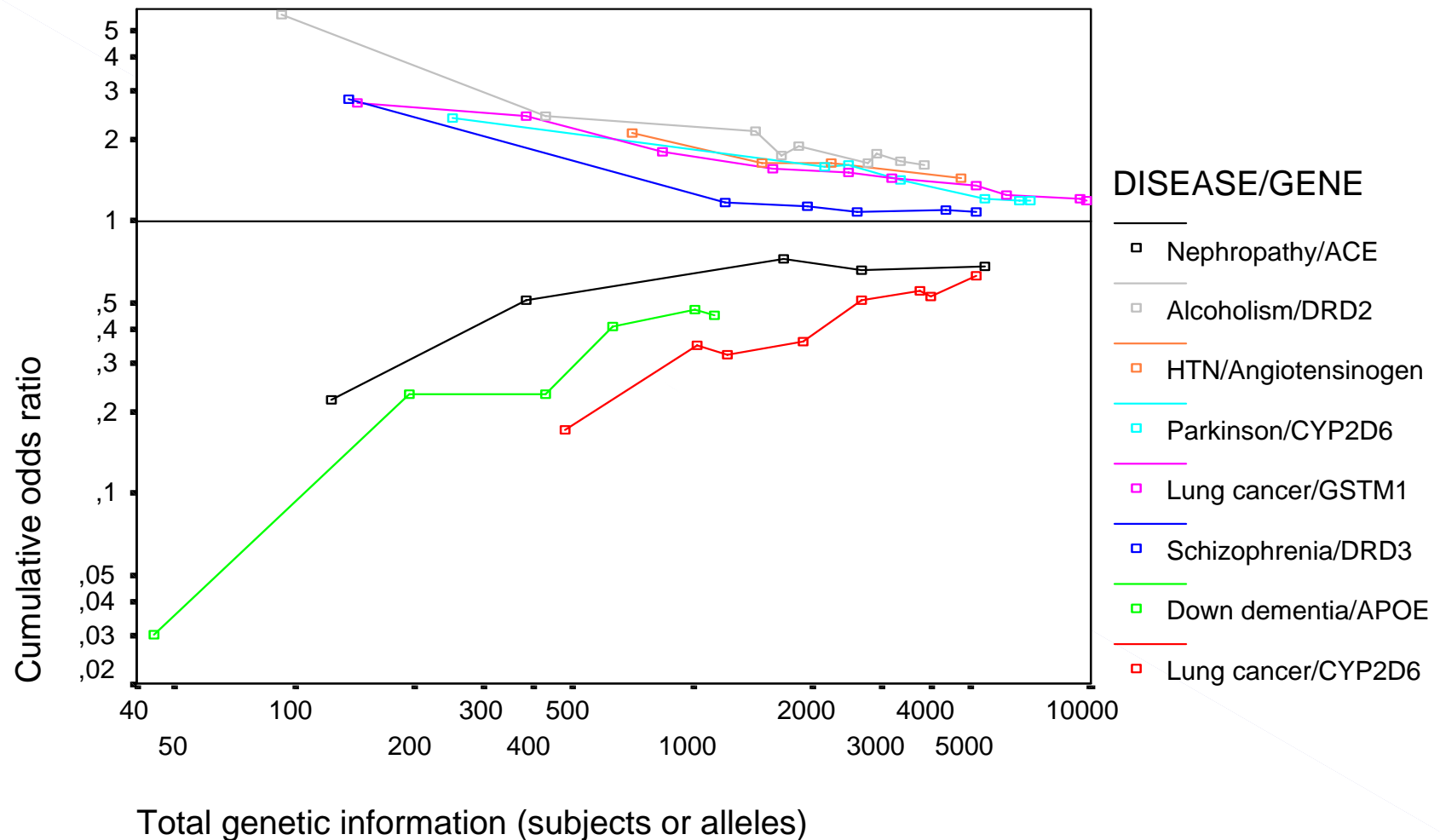
Counting fish in the sea of association analyses

Multiplier	Parameter
>100000000	Gene variants
>1000	Diseases
>10	Outcomes
>10	Subgroups
>10	Genetic contrasts
>10	Investigators
1 quadrillion	Candidate analyses

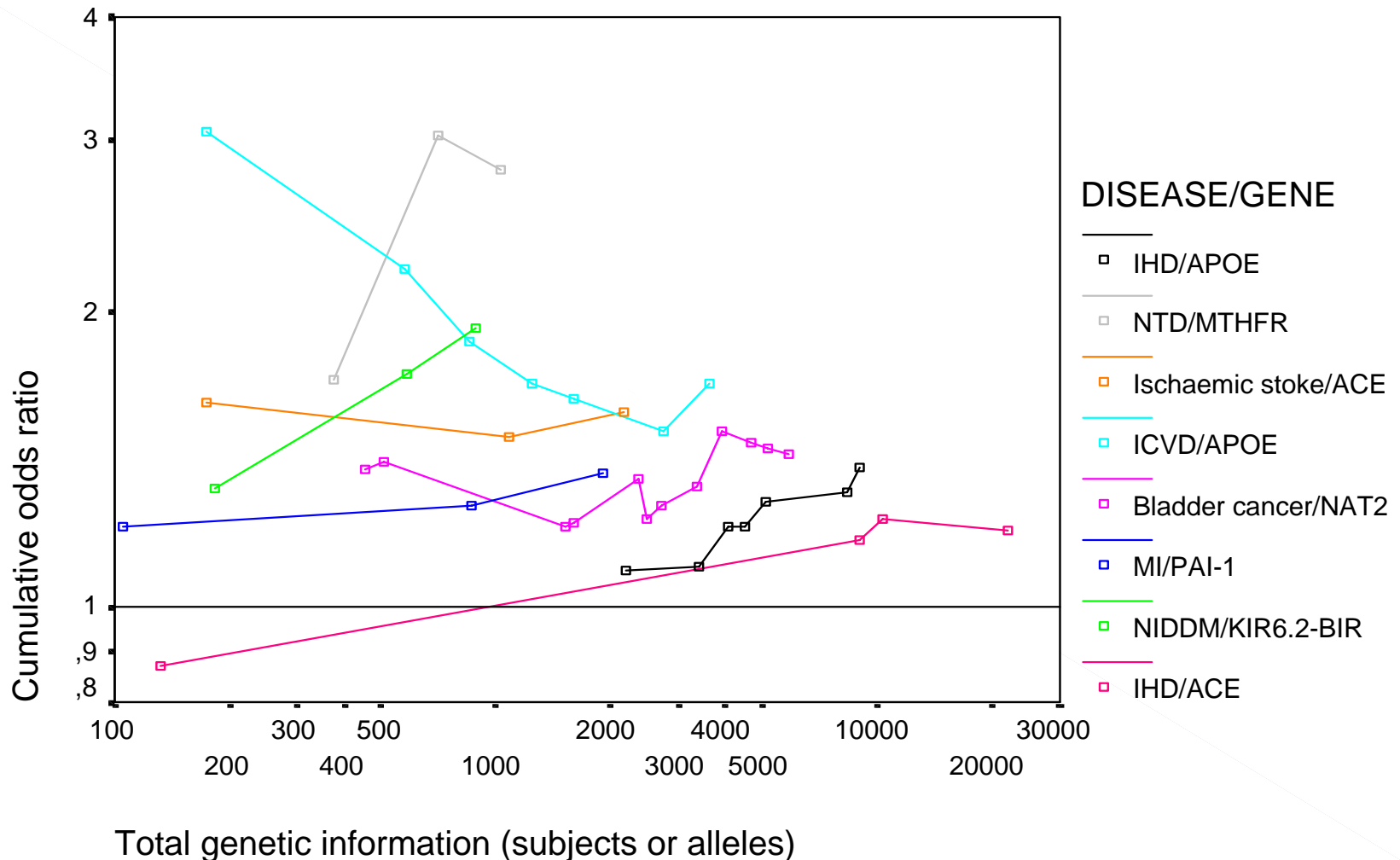
The legend of focusing “based on biological plausibility”

- Just in the year 2002 studies were published addressing the relationship of the APOE epsilon polymorphism with familial Alzheimer’s disease; sporadic Alzheimer’s disease; colorectal cancer; fatty liver; atherosclerosis; hyperlipidemia; acute ischemic stroke; spina bifida; coronary artery disease; normal tension glaucoma; hypertension; Parkinson’s disease, diabetic nephropathy; pre-eclampsia; hepatitic C-related liver disease; cerebrovascular disease; coronary artery disease post-renal transplantation; non-specified cognitive impairment; childhood nephrotic syndrome; spontaneous abortion; multiple sclerosis; alcohol withdrawal; cognitive dysfunction after coronary artery surgery; alcoholic chronic pancreatitis; alcoholic cirrhosis; macular toxicity from chloroquine; macular edema; aortic valve stenosis; vascular dementia; type II diabetes mellitus; and migraine.

Evolving effect sizes: spurious effects that diminish/disappear over time



Effects that are not significant originally, but become so eventually



Predictors of statistically significant discrepancies between the first and subsequent studies on the same genetic association.

Predictor of discrepancy	Univariate regressions	
	OR (95% CI)	<i>P</i> -value
Total number of studies (per study)	1.17 (1.03-1.33)	.020
Sample size of first study(ies) (doubling)	0.42 (0.17-0.98)	.046
Single first study with clear genetic contrast	9.33 (1.01-86.3)	.044

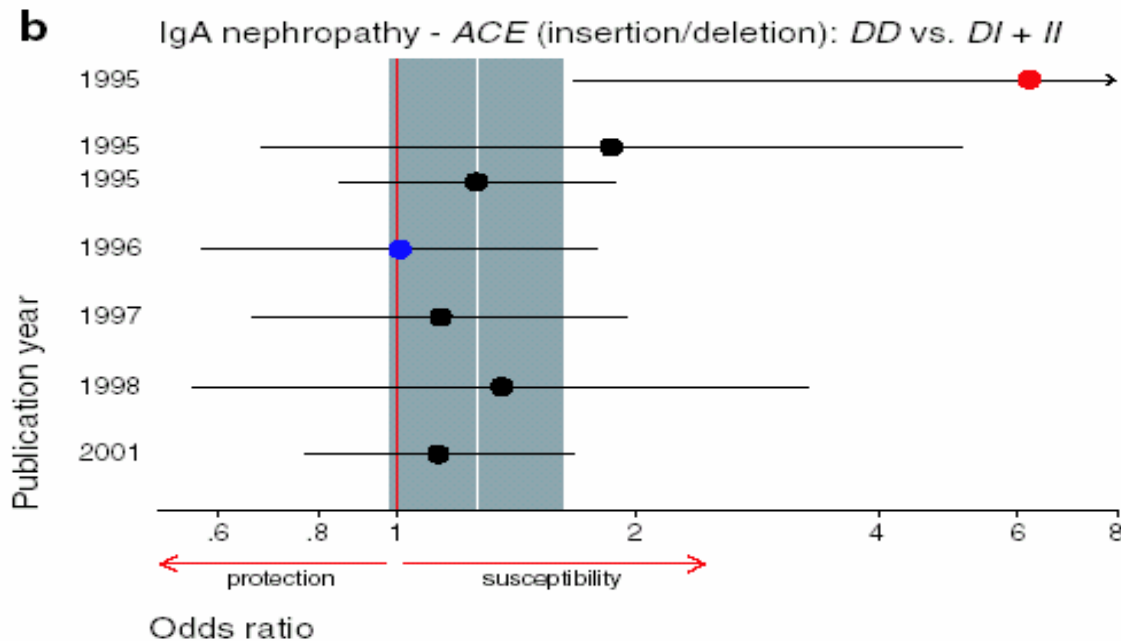
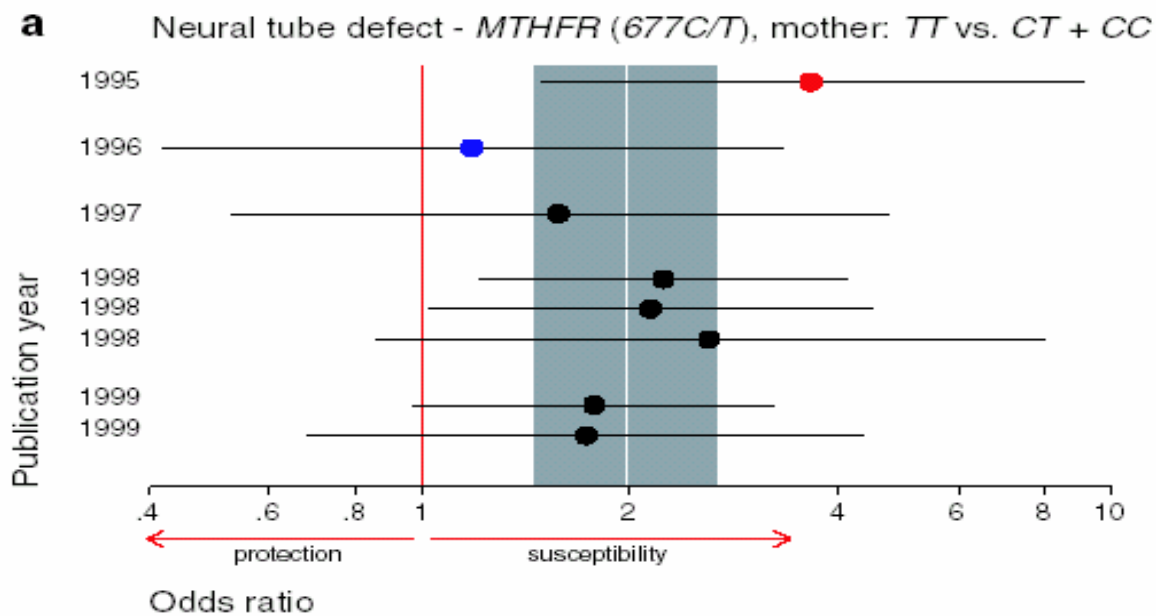
Large vs. small studies

- They offer give different results and the more usual scenario is that large studies give more conservative or null results
- Publication bias?
- Hints of other reporting biases?
- Genuine heterogeneity?

m - a	H	R	F	D 1	D 2	D 3	R S	F S
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								
12								
13								
14								
15								
16								
17								
18								
19								
20								
21								
22								
23								
24								
25								
26								
27								
28								
29								
30								
31								
32								
33								
34								
35								
36								
37								
38								
39								
40								
41								
42								
43								
44								
45								
46								
47								
48								
49								
50								
51								
52								
53								
54								
55								

H: heterogeneity
 R/F: difference in first vs.
 subsequent
 D1-D3: publication bias
 diagnostics
 RS/FS: significant findings
 (with/without first studies)

Ioannidis et al, Lancet 2003



Succession of
early
extremes:
Proteus
phenomenon

Ioannidis et al (in press)

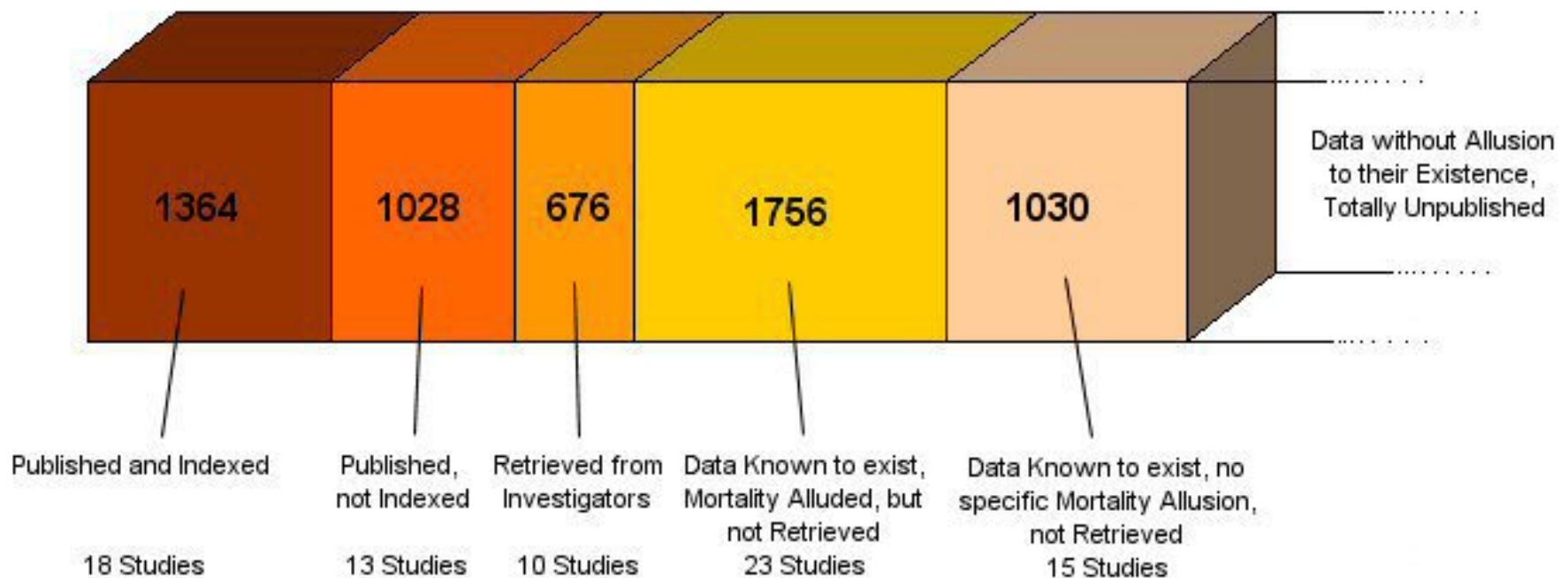
Racial (or other subgroup) differences?

- Empirical evidence suggest that while allele frequencies differ a lot ($I\text{-squared} \geq 75\%$) in 58% of postulated gene-disease associations, differences in the effect sizes (odds ratios) occur in 14%.
- No differences in race-specific odds ratios have been recorded once we have exceeded a total sample size of $N=10,000$

Problems of standardization

- Polymorphic markers
- Quantitative traits, intermediate/surrogate endpoints
- Time-dependent effects
- Too much analytical liberty

Readily available, available, hidden, and very well hidden data: a real example on a prognostic factor for survival



Options for integration of information

- Single, all-absorbing mega-studies (e.g. proposed US cohort on genes and environment)
- Meta-analyses of group data
- Meta-analyses of individual participant data
- All of these designs are unlikely to be successful unless they allow for evolving (often rapidly evolving) evidence

Advantages of MIPD

Advantages

Data

- More information

- Inclusion of extended databases from published studies

- Inclusion of data from unpublished studies

Better standardization of information

- Categorization of eligible participants

- Outcomes

- Definition of follow-up period and censoring criteria

Analysis

- Better time-to-event analyses

- Standardized statistical models

- Evaluation of time-dependency

Better adjusted/multivariate analyses

- Consistent treatment of loci in linkage disequilibrium

- Evaluation of dose-response effects for multiple genes or double doses of a single gene

- Evaluation of subgroup effects, including racial heterogeneity

Interpretation

- Assessment of heterogeneity

- Assessment of sampling bias in specific studies

Other

- Establishment of international networks of collaborating investigators

Disadvantages of MIPD

Data

Data may not be made available from all published studies

Interpretation

Potential post hoc conflicts with collaborators regarding findings

Resources

Substantial effort and infrastructure required to:

- Develop and administer a standardized protocol

- Collect, manage, and analyze data

- Communicate with collaborators

Study registration

- As of the fall of 2004, most major medical journals have agreed that they will not publish any randomized trials unless they are registered in an accredited trial registry when they are initiated
- This is expected to increase transparency, and reduce selection biases in clinical research
- Can this be done for molecular medicine: can one register upfront all a priori hypotheses – especially in public? This would be counterintuitive to the competitive “discovery” spirit of basic research.

An alternative: investigator or data specimen registration

- Inclusive networks of investigators working on the same disease, set of genes or field
- Promotion of better methods and standardization
- Research freedom for individual participating teams
- Thorough and unbiased testing of proposed hypotheses with promising preliminary data on large-scale comprehensive databases
- Due credit to investigators for both “positive” and “negative” findings
- It is feasible to start from existing coalitions of investigators (“networks”) that work on specific diseases, genes or fields

Registries of teams

- The core registry should comprise information on the teams that already participate in a network
- A wider registry should also record all other teams that work on the same field. This should be based on searches of electronic databases (identifying who has published anything on the field of interest), personal contacts, announcement in some major journal (e.g. commentary currently in peer review) and should be an open, evolving process updated at regular intervals
- Depending on the structure and funding opportunities of the existing networks, additional teams may be allowed to join formally and fully in the original network; even if structure or funding considerations do not allow this, additional teams should be simply recorded, so that a picture of the field-at-large is available
- Networks may have qualitative or other pre-requisites for allowing teams to join. These should be developed by the scientists involved, but some central guidance and sharing of experiences would also be useful

How might it look like?

- For cancer X, a network is available with 43 participating teams and with a total of 25000 cases and 27000 controls (total 52000)
- Besides the network, we are also aware of the existence of another 28 teams working on the genetics of this cancer with a total of 18000 cases and 17000 controls (total 35000)
- Promising findings from single teams or findings from meta-analyses of published group data may be tested on a large-scale at the network level
- The certainty for any preliminary finding can be interpreted not only as a function of its statistical significance, but also as a function of the percentage of the total possible evidence upon which it is based; e.g. an odds ratio may have a p-value of 0.001 after 4 teams have tested a specific SNP, but this may be based only on 2600 subjects, i.e. 5% of the total network possible evidence and approximately 3% of the overall possible evidence.
- The network would also ensure that “negative” findings are also disseminated with appropriate credit

Examples of investigator networks: disease-specific

- GENOMOS (osteoporosis)
- GEO-PD (Parkinson's disease)
- Interlymph (lymphoma)
- ILCCO (Lung cancer)
- INHANCE (head and neck cancer)
- Meta-analysis of HIV Host Genetics (HIV)
- WHO craniofacial anomalies consortium (craniofacial anomalies)
- Emerging Risk Factors Collaboration (cardiovascular disease)

Examples of networks: gene- or field-specific

- GSEC (genes involved in environmental carcinogens)
- Web registry of DNA repair genes and cancer
- US Pharmacogenetics Research Network

What would a network of networks do

- Communication and sharing of expertise in statistical analytical methods, laboratory techniques, practical procedures, logistics of creating and maintaining a network
- Co-ordination of registries, facilitation and avoidance of overlap
- Maximization of efficiency and standardization of methods and procedures
- Electronic list of all registries containing minimal information on all participating teams as well as on non-participating teams
- Eventually keeping updated a “Libro d’oro” of validated molecular information that may be compiled by investigators of each network for the disease/genes/field-at hand

Eventual proposed grading of evidence in molecular research

- III. Single or scattered studies: purely hypothesis-generating, important to register data, regardless of results
- II. Meta-analyses of group data: increasing certainty when several thousand subjects available
- I. Large-scale evidence from individual-level all-inclusive networks: evolving gold standard?

- C. No functional/biological data or negative data
- B. Limited or controversial functional data
- A. Convincing functional data

- 3. No clinical or public health applicability
- 2. Limited applicability
- 1. Clinical/public health applicability